Introduction to Protein Structure

Second Edition

Carl Branden

Microbiology and Tumor Biology Center
Karolinska Institute
Stockholm
Sweden

John Tooze

Imperial Cancer Research Fund Laboratories
Lincolns Inn Fields
London
UK

of tertiary structure is still slow compared with the rate of accumulation of amino acid sequence data. This makes the **folding problem**, the successful prediction of a protein's tertiary structure from its amino acid sequence, central to rapid progress in post-genomic biology. We will, therefore, in this chapter first briefly describe implications of protein homology and methods for the prediction of secondary and tertiary structure before giving some examples of protein engineering and protein design.

Homologous proteins have similar structure and function

The term **homology** as used in a biological context is defined as similarity of structure, physiology, development and evolution of organisms based upon common genetic factors. The statement that two proteins are homologous therefore implies that their genes have evolved from a common ancestral gene.

Homologous proteins are mostly recognized by statistically significant similarities in their amino acid sequences. Usually, they also have similar functions although there are some known exceptions, where genes for ancient enzymes have been recruited at a later stage in evolution to produce proteins with quite different functions. An example is provided by one of the structural components in the eye lens that is homologous to the ancient glycolytic enzyme lactate dehydrogenase. Once a novel gene has been cloned and sequenced, a search for amino acid sequence similarity between the corresponding protein and other known protein sequences should be made. Usually, this is done by comparison with databases of known protein sequences using one of the standard sequence alignment computer programs.

Two proteins are considered to be homologous when they have identical amino acid residues in a significant number of sequential positions along the polypeptide chains. Using statistical methods based on comparisons of computer-generated random sequences, it is relatively straightforward to assess how many positions need to be identical for a statistically significant identity between two sequences. However, it is frequently found that two proteins with sequence identity below the level of statistical significance have similar functions and similar three-dimensional structures. In these cases, functionally important residues are identical and usually such residues form sequence patterns or motifs that can be used to identify other proteins that belong to the same functional family. Frequently, members of such families are also considered to be homologous, even though the identities are not statistically significant, only functionally significant. Databases for such families, based on identical or similar sequence motifs, are available on the World Wide Web (see pp. 393–394) and they are very useful for assigning function to a novel protein.

If significant amino acid sequence identity is found with a protein of known crystal structure, a three-dimensional model of the novel protein can be constructed, using computer modeling, on the basis of the sequence alignment and the known three-dimensional structure. This model can then serve as an excellent basis for identifying amino acid residues involved in the active site or in antigenic epitopes, and the model can be used for protein engineering, drug design, or immunological studies.

Since the sequence databases are large and growing exponentially, currently comprising more than 500,000 known protein sequences, the standard sequence alignment programs have been designed to provide a compromise between the speed and the accuracy of the search. As a result, they work well only when there is a reasonably high degree of sequence identity, usually of the order of 30% or more. Much more sensitive programs have been written that search for both identity and conserved structural properties and also for relatedness in different physical properties, but these inevitably require far more computing time. Carefully used, such programs can identify structural and functional similarity where the standard programs fail to do so.

Homologous proteins have conserved structural cores and variable loop regions

Homologous proteins always contain a core region where the general folds of the polypeptide chains are very similar. This core region contains mainly the secondary structure elements that build up the interior of the protein: in other words, the scaffolds of homologous proteins have similar three-dimensional structures. Even distantly related proteins with low sequence identity have similar scaffold structures, although minor adjustments occur in the positions of the secondary structure elements to accommodate differences in the arrangements of the hydrophobic side chains in the interior of the protein. The greater the sequence identity, the more closely related are the scaffold structures (Figure 17.1). This has important implications for model building of homologous proteins; the more distantly related two proteins are, the more the scaffold must be adjusted to model the new structure.

Loop regions that connect the building blocks of scaffolds can vary considerably both in length and in structure. The problem of predicting the three-dimensional structure of a protein that is homologous to a protein of known three-dimensional structure is therefore mainly a question of predicting the structure of loop regions and side-chain conformations, after the scaffold has been adjusted. As mentioned in Chapter 2, loop regions do not have random structures, and their main-chain conformations cluster in sets of similar structures. The conformation of each set depends more on the number of amino acids in the loop and the type of secondary structure elements that it connects, whether they are α-α, β-β, α-β, or β-α connections, than on the actual amino acid sequences. Therefore it is possible to use a database of loop regions from proteins of known structure to obtain a preliminary model of the loops of an unknown structure. To model a protein structure, suitable main-chain loop conformations from this database are attached to the scaffold modeled to have a structure similar to that of the known homologous protein. Finally, the conformations of the side chains are predicted by energy refinement of the model, which minimizes the free energy of the protein by maximizing the interaction energies of the amino acids. Analysis of structures determined to high resolution has shown that only a few side-chain conformations frequently occur. These are called rotamers and model building of side chains employs databases of such rotamers.

An instructive example of the use of such procedures has been in modeling antigen-binding sites in immunoglobulins. These binding sites are built up from three hypervariable loop regions, CDR1–CDR3, from the variable domains of both the light and the heavy chains of immunoglobulins as described in Chapter 15. There is usually high sequence identity within the scaffolds of the variable domains in different immunoglobulin molecules. Consequently, the scaffold of variable domains of known three-dimensional structures can be used in modeling a new monoclonal antibody with a known amino acid sequence. However, the CDR regions of a new antibody are usually very different in sequence from those of any other known antibody, and their three-dimensional structures must be predicted. By comparing

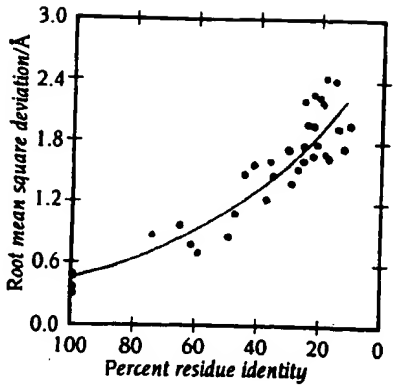
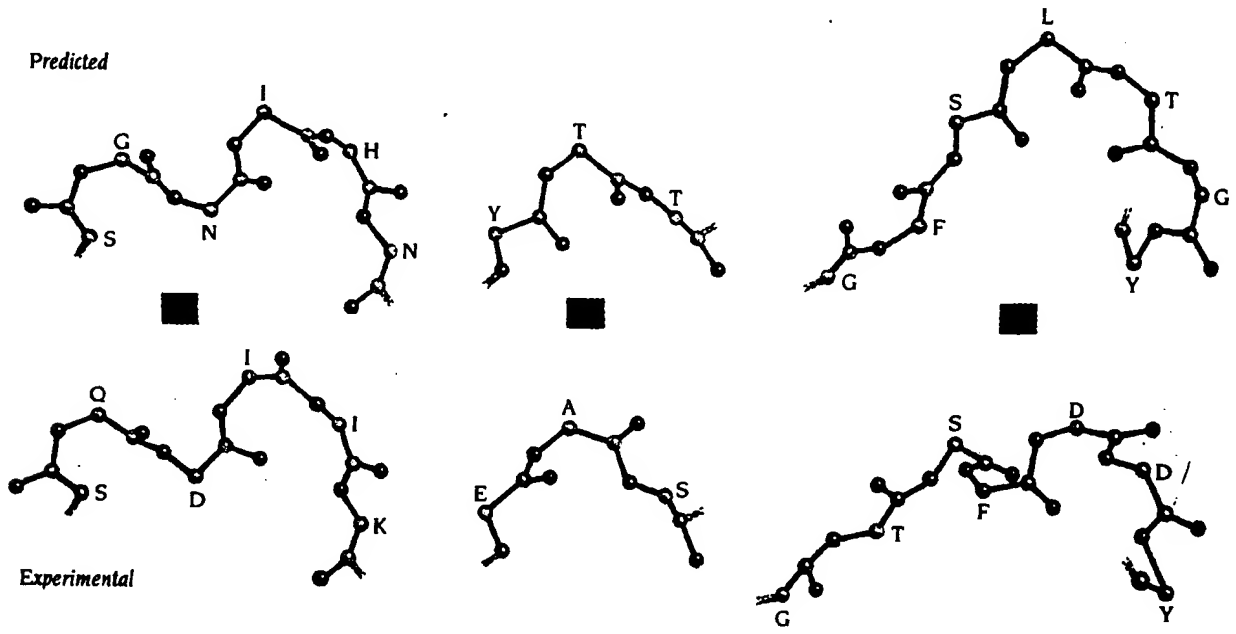


Figure 17.1 The relation between the divergence of amino acid sequence and three-dimensional structure of the core region of homologous proteins. Known structures of 32 pairs of homologous proteins such as globins, serine proteinases, and immunoglobulin domains have been compared. The root mean square deviation of the main-chain atoms of the core regions is plotted as a function of amino acid homology (red dots). The curve represents the best fit of the dots to an exponential function. Pairs with high sequence homology are almost identical in three-dimensional structure, whereas deviations in atomic positions for pairs of low homology are of the order of 2 Å. (From C. Chothia and A. Lesk, *EMBO J.* 5: 823–826, 1986.)



known antibody structures and sequences, it has been shown that there is only a small repertoire of main-chain conformations for at least five of the six CDR regions and that the particular conformation adopted is determined by a few key conserved residues for each loop conformation. For example, three different conformations were found for the CDR3 regions of the light chains in nine known x-ray structures. More than 90% of the known sequences of light-chain CDR3 regions obey the sequence constraints of one or other of these three conformations. By using this repertoire of loop conformations, considerable success has been achieved in correctly predicting the structure of antigen-binding surfaces. An example of such a prediction compared with the actual structure, subsequently determined, is given in Figure 17.2.

Figure 17.2 An example of prediction of the conformations of three CDR regions of a monoclonal antibody (*top row*) compared with the unrefined x-ray structure (*bottom row*). L1 and L2 are CDR regions of the light chain, and H1 is from the heavy chain. The amino acid sequences of the loop regions were modeled by comparison with the sequences of loop regions selected from a database of known antibody structures. The three-dimensional structure of two of the loop regions, L1 and L2, were in good agreement with the preliminary x-ray structure, whereas H1 was not. However, during later refinement of the x-ray structure errors were found in the conformations of H1, and in the refined x-ray structure this loop was found to agree with the predicted conformations. In fact, all six loop conformations were correctly predicted in this case. (From C. Chothia et al., *Science* 233: 755–758, 1986.)

Knowledge of secondary structure is necessary for prediction of tertiary structure

What can be done by predictive methods if the sequence search fails to reveal any homology with a protein of known tertiary structure? Is it possible to model a tertiary structure from the amino acid sequence alone? There are no methods available today to do this and obtain a model detailed enough to be of any use, for example, in drug design and protein engineering. This is, however, a very active area of research and quite promising results are being obtained; in some cases it is possible to predict correctly the type of protein, α, β, or α/β, and even to derive approximations to the correct fold.

Today's predictive methods rely on prediction of secondary structure: in other words, which amino acid residues are α-helical and which are in β strands. We have emphasized in Chapter 12 that secondary structure cannot in general be predicted with a high degree of confidence with the possible exceptions of transmembrane helices and α-helical coiled coils. This imposes a basic limitation on the prediction of tertiary structure. Once the correct secondary structure is known, we know enough about the rules for packing elements of secondary structure against each other (see Chapter 2 for helix packing) to derive a very limited number of possible stable globular folds. Consequently, secondary structure prediction lies at the heart of the prediction of tertiary structure from the amino acid sequence.